Submitted to the Annals of Statistics

CRITERIA FOR BAYESIAN MODEL CHOICE WITH APPLICATION TO VARIABLE SELECTION*

By M.J. Bayarri†, J.O. Berger‡ A. Forte§ and G. García-Donato¶

In objective Bayesian model selection, no single criterion has emerged as dominant in defining objective prior distributions. Indeed, many criteria have been separately proposed and utilized to propose differing prior choices. We first formalize the most general and compelling of the various criteria that have been suggested, together with a new criterion. We then illustrate the potential of these criteria in determining objective model selection priors by considering their application to the problem of variable selection in normal linear models. This results in a new model selection objective prior with a number of compelling properties.

1. Introduction.

1.1. *Background.* A key feature of Bayesian model selection, when the models have differing dimensions and non-common parameters, is that results are typically highly sensitive to the choice of priors for the non-common parameters and, unlike the scenario for estimation, this sensitivity does not vanish as the sample size grows (see Kass and Raftery, 1995; Berger and Pericchi, 2001). Furthermore, improper priors cannot typically be used for non-common parameters, nor can 'vague proper priors' (see the above references, for example, and the brief discussion in Section 2.2), ruling out use of the main tools developed in objective Bayesian estimation theory.

*This paper was supported in part by the Spanish Ministry of Education and Science under grant MTM2010-19528, and by USA National Science Foundation Grants DMS-0635449, DMS-0757549-001, and DMS-1007773.

†Universitat de València, Valencia, Valencia, Spain.

‡Duke University, Durham, North Carolina, USA.

§Universitat Jaume I, Castellón, Valencia, Spain

¶Universidad de Castilla-La Mancha, Albacete, Castilla-La Mancha, Spain

AMS 2000 subject classifications: Primary 62J05, 62J15; secondary 62C10

Keywords and phrases: Model Selection, Variable Selection, Objective Bayes

Because of the difficulty in assessing subjective priors for numerous models, there have been many efforts (over more than 30 years) to develop 'conventional' or 'objective' priors for model selection; we will term these 'objective model selection priors,' the word objective simply meant to indicate that they are not subjective priors, and are chosen conventionally based on the models being considered. A few of the many references most related to this paper are Jeffreys (1961); Zellner and Siow (1980, 1984); Laud and Ibrahim (1995); Kass and Wasserman (1995); Berger and Pericchi (1996); Moreno, Bertolino and Racugno (1998); De Santis and Spezzaferri (1999); Pérez and Berger (2002); Bayarri and García-Donato (2008); Liang et al. (2008); Cui and George (2008); Maruyama and George (2008); Maruyama and Strawderman (2010).

For the most part, these efforts have started with a good idea, used it to develop the priors, and then studied the behavior of the priors. Yet, in spite of the apparent success of many of these methods, there has been no agreement as to which are most appealing or most successful.

This lack of progress in reaching consensus on objective priors for model selection resulted in our approaching the problem from a different direction, namely formally formulating the various criteria that have been deemed essential for model selection priors (such as consistency of the resulting procedure), and seeing if these criteria can essentially determine the priors.

The criteria are stated for general model selection problems in Section 2, which also discusses their historical antecedents. To illustrate that application of the criteria can largely determine model selection priors, we turn to a specific problem in Section 3 – variable selection in normal linear models. The resulting priors for variable selection are new and result in closed form Bayes factors; for those primarily interested in the methodology itself, the resulting priors and Bayes factors are given in Section 4.

1.2. *Notation.* Let $\boldsymbol{y}$ be a data vector of size n from one of the models

$$M_0 : f_0(\boldsymbol{y} \mid \boldsymbol{\alpha}), \quad M_i : f_i(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_i), \;\; i = 1, 2, \ldots, N-1\,, \tag{1}$$

where $\boldsymbol{\alpha}$ and the $\boldsymbol{\beta}_i$ are unknown model parameters, the latter having dimension k_i. M_0 will be called the null model and is nested in all of the

considered models.

Under the null model, the prior is $\pi_0(\boldsymbol{\alpha})$; under model M_i, and without loss of generality, we express the model selection prior as

$$\pi_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_i) = \pi_i(\boldsymbol{\alpha})\, \pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha}).$$

Note that the parameter $\boldsymbol{\alpha}$ occurs in all of the models, so that $\boldsymbol{\alpha}$ is usually referred to as the *common* parameters; the $\boldsymbol{\beta}_i$ are called *model specific* parameters.

Assuming that one of the entertained models is true, the posterior probability of each of the models M_i can be written in the convenient form

$$Pr(M_i \mid \boldsymbol{y}) = \frac{B_{i0}}{1 + \left(\sum_{j=1}^{N-1} B_{j0}\, P_{j0}\right)}\,, \tag{2}$$

where P_{j0} is the prior odds $P_{j0} = Pr(M_j)/Pr(M_0)$, with $Pr(M_j)$ being the prior probability of model M_j, and B_{j0} is the Bayes factor of model M_j to M_0 defined by

$$B_{j0} = \frac{m_j(\boldsymbol{y})}{m_0(\boldsymbol{y})}, \text{ with } m_j(\boldsymbol{y}) = \int f_j(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_i)\, \pi_j(\boldsymbol{\alpha},\, \boldsymbol{\beta}_j)\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta}_j \tag{3}$$

and $m_0(\boldsymbol{y}) = \int f_0(\boldsymbol{y} \mid \boldsymbol{\alpha})\pi_0(\boldsymbol{\alpha}) d\boldsymbol{\alpha}$ being the marginal likelihoods of model M_j and M_0 corresponding to the model prior densities $\pi_j(\boldsymbol{\alpha},\, \boldsymbol{\beta}_j)$ and $\pi_0(\boldsymbol{\alpha})$. (Any model could serve as the base model for computation of the Bayes factors in (2), but use of the null model is common and convenient.) The focus in this paper is on choice of model priors $\pi_0(\boldsymbol{\alpha})$ and $\pi_j(\boldsymbol{\alpha},\, \boldsymbol{\beta}_j)$.

2. Criteria for objective model selection priors.

2.1. *Introduction.* The arguments concerning prior choice in testing and model selection in Jeffreys (1961) are often called Jeffreys' *desiderata* (see Berger and Pericchi, 2001) and are the precursors to the criteria developed herein. (Robert, Chopin and Rousseau, 2009, is a comprehensive and modern review of Jeffreys' book.) These and related ideas have been repeatedly used to evaluate or guide development of objective model priors (see e.g. Berger and Pericchi, 2001; Bayarri and García-Donato, 2008; Liang et al., 2008; and Forte, 2011). We group the criteria into four classes: basic, consistency criteria, predictive matching criteria, and invariance criteria.

2.2. *Basic criteria.* As mentioned in the Introduction, priors for the non-common parameters $\boldsymbol{\beta}_i$ should be proper, because they only occur in the numerator of the Bayes factors B_{i0} and hence, if using an improper prior, the arbitrary constant for the improper prior would not cancel, making B_{i0} ill-defined. There have been various efforts at using improper priors and defining a meaningful scaling (Ghosh and Samanta, 2002; Spiegelhalter and Smith, 1982); and other methods have been proposed that can be interpreted as implicitly scaling the improper prior Bayes factor (see details and references in Bayarri and García-Donato, 2008), but we are restricting consideration here to real Bayesian procedures.

Similarly, vague proper priors cannot be used in determining the B_{i0}, since the arbitrary scale of vagueness appears as a multiplicative term in the Bayes factor, again rendering the Bayes factor arbitrary. Thus we have

Criterion 1 - Basic: *Each conditional prior $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})$ must be proper (integrating to one) and cannot be arbitrarily vague in the sense of almost all of its mass being outside any believable compact set.*

2.3. *Consistency criteria.* Following Liang et al. (2008), we consider two primary consistency criteria – model selection consistency and information consistency:

Criterion 2 - Model selection consistency: *If data $\boldsymbol{y}$ have been generated by M_i, then the posterior probability of M_i should converge to 1 as the sample size $n \to \infty$.*

Model selection consistency is not particularly controversial, although it can be argued that the true model is never one of the entertained models, so that the criterion is vacuous. Still, it would be philosophically troubling to be in a situation with infinite data generated from one of the models being considered, and not choosing the correct model. A number of recent references concerning this criterion are Fernández, Ley and Steel (2001); Berger, Ghosh and Mukhopadhyay (2003); Liang et al. (2008); Casella et al. (2009); Guo and Speckman (2009).

Criterion 3 - Information consistency: *For any model M_i, if $\{\boldsymbol{y}_m, m =$*

$1, \ldots\}$ *is a sequence of data vectors of fixed size such that, as* $m \to \infty$,

$$\Lambda_{i0}(\boldsymbol{y}_m) = \frac{\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}_i} f_i(\boldsymbol{y}_m \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_i)}{\sup_{\boldsymbol{\alpha}} f_0(\boldsymbol{y}_m \mid \boldsymbol{\alpha})} \to \infty\,, \quad \textit{then} \quad B_{i0}(\boldsymbol{y}_m) \to \infty\,. \tag{4}$$

In normal linear models, this is equivalent to saying that, if one considers a sequence of data vectors for which the corresponding F (or t) statistic goes to infinity, then the Bayes factor should also do so for this sequence. Jeffreys (1961) used this argument to justify a Cauchy prior in testing that a normal mean is zero, and the argument has also been highlighted in Berger and Pericchi (2001); Bayarri and García-Donato (2008); Liang et al. (2008). One can construct examples in which a real Bayesian answer violates information consistency, but the examples are based on very small sample sizes and priors with extremely flat tails. Furthermore, violation of information consistency would place frequentists and Bayesians in a particularly troubling conflict, which many would view as unattractive.

A third type of consistency has been proposed to address the fact that objective model selection priors typically depend on specific features of the model, such as the sample size or the particular covariates being considered.

Criterion 4 - Intrinsic prior consistency: *Let* $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha}, n)$ *denote the prior for the model specific parameters of model* M_i *with sample size* n*. Then, as* $n \to \infty$ *and under suitable conditions on the evolution of the model with* n, $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha}, n)$ *should converge to a proper prior* $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})$.

The idea here is that, while features of the model, sample size (and possibly even data) frequently affect model selection priors, such features should disappear for large n. If there is such a limiting prior it is called an *intrinsic prior*; see Berger and Pericchi (2001) for extensive discussion and previous references. (Note that some have used the phrase 'intrinsic prior' to refer to specific priors arising from a specific model selection method, but we use the term here generically.)

2.4. *Predictive matching criteria.* The most crucial aspect of objective model selection priors is that they be appropriately 'matched' across models of different dimensions. Having a prior scale factor 'wrong' by a factor of 2

does not matter much in one dimension, but in 50 dimensions that becomes an error of 2^{50} in the Bayes factor. There have been many efforts to achieve such matching in model selection, including Spiegelhalter and Smith (1982); Suzuki (1983); Laud and Ibrahim (1995); Ghosh and Samanta (2002).

The standard approach to predictive matching is modeled after Jeffreys (1961). For example, Jeffreys defined a 'minimal sample size' for which one would logically be unable to discriminate between two hypotheses and argued that the prior distributions should be chosen to then yield equal marginal likelihoods for the two hypotheses. Here is an illustration of this type of argument, from Berger, Pericchi and Varshavsky (1998).

Example: Suppose one is comparing two location-scale models

$$M_1 : y \sim \frac{1}{\sigma}\, p_1\left(\frac{y-\mu}{\sigma}\right) \quad \text{and} \quad M_2 : y \sim \frac{1}{\sigma}\, p_2\left(\frac{y-\mu}{\sigma}\right).$$

Intuitively, two independent observations (y_1, y_2) should not allow for discrimination between these models, since two observations only allow setting of the center and scale of the distribution; there are no 'degrees of freedom' left for model discrimination. Now consider the choice of prior (for both models) $\pi(\mu, \sigma) = 1/\sigma$. It is shown in Berger, Pericchi and Varshavsky (1998) that

$$\int \frac{1}{\sigma^2} p_1\left(\frac{y_1-\mu}{\sigma}\right) p_1\left(\frac{y_2-\mu}{\sigma}\right) \pi(\mu,\sigma) d\mu d\sigma$$
$$= \int \frac{1}{\sigma^2} p_2\left(\frac{y_1-\mu}{\sigma}\right) p_2\left(\frac{y_2-\mu}{\sigma}\right) \pi(\mu,\sigma)\; d\mu d\sigma = \frac{1}{2|y_1-y_2|},$$

for any pair of observations $y_1 \neq y_2$, so that the models would be said to be predictively matched for all minimal samples. The Bayes factor between the models is then obviously 1, agreeing with the earlier intuition that a minimal sample should not allow for model discrimination.

This argument was formalized by Berger and Pericchi (2001) as follows.

DEFINITION 1. *The model/prior pairs* $\{M_i, \pi_i\}$ *and* $\{M_j, \pi_j\}$ *are* predictive matching *at sample size* n^* *if the predictive distributions* $m_i(\boldsymbol{y}^*)$ *and* $m_j(\boldsymbol{y}^*)$ *are close in terms of some distance measure for data of that sample size. The model/prior pairs* $\{M_i, \pi_i\}$ *and* $\{M_j, \pi_j\}$ *are* exact predictive matching *at sample size* n^* *if* $m_i(\boldsymbol{y}^*) = m_j(\boldsymbol{y}^*)$ *for all* $\boldsymbol{y}^*$ *of sample size* n^*.

One only wants predictive matching for 'minimal' sample sizes, since, for larger sample sizes, the discrimination between models occurs through the marginal densities; they must differ for discrimination.

Criterion 5 - Predictive matching: *For appropriately defined 'minimal sample size' in comparing M_i with M_j, one should have model selection priors that are predictive matching. Optimal (though not always obtainable) is exact predictive matching.*

In Berger and Pericchi (2001), minimal sample size was defined as the smallest sample size for which the models under consideration have finite marginal densities when objective estimation priors are used. Typically this minimal sample size equals the number of parameters in the model or, more generally, is the number of observations needed for all parameters to be identifiable. For model selection, however, minimal sample size needs to be defined relative to the model selection priors being utilized. Hence we have the following general definition.

DEFINITION 2 (**Minimal training sample**). *A* minimal training sample $\boldsymbol{y}_i^*$ *for* $\{M_i, \pi_i\}$ *is a sample of minimal size* $n_i^* \geq 1$ *with a finite non-zero marginal density* $m_i(\boldsymbol{y}_i^*)$.

There are many possibilities for even exact predictive matching. We here highlight two types of exact predictive matching, which are of particular relevance to the development of objective model selection priors for the variable selection problem discussed in Section 3.

DEFINITION 3 (**Null predictive matching**). *The model selection priors are* null predictive matching *if each of the model/prior pairs* $\{M_i, \pi_i\}$ *and* $\{M_0, \pi_0\}$ *are exact predictive matching for all minimal training samples* $\boldsymbol{y}_i^*$ *for* $\{M_i, \pi_i\}$.

Definition 3 reflects the common view – starting with Jeffreys (1961) – that data of a minimal size should not allow one to distinguish between the null and alternative models. Null predictive matching arguments have also been used by Ghosh and Samanta (2002) and Spiegelhalter and Smith (1982) among others.

DEFINITION 4 (**Dimensional predictive matching**). *The model selection priors are* dimensional predictive matching *if each of the model/prior pairs* $\{M_i, \pi_i\}$ *and* $\{M_j, \pi_j\}$ *of the same complexity/dimension (i.e.* $k_i = k_j$*) are exact predictive matching for all minimal training samples* $\boldsymbol{y}_i^*$ *for models of that dimension.*

The next section gives the most prominent example of dimensional predictive matching.

2.5. *Invariance criteria.* Invariance arguments have played a prominent role in statistics (cf. Berger, 1985), especially in objective Bayesian estimation theory. They are also extremely helpful in part of the specification of objective Bayesian model selection priors.

A basic type of invariance that is almost always relevant for model selection is invariance to the units of measurement being used:

Criterion 6 - Measurement invariance: *The units of measurement used for the observations or model parameters should not affect Bayesian answers.*

A much more powerful, but special, type of invariance arises when the family of models under consideration are such that the model structures are invariant to group transformations. Following the notation in Berger (1985), we formally state

DEFINITION 5. *The family of densities for* $\boldsymbol{y} \in \mathbb{R}^n$, $\mathfrak{F} := \{f(\boldsymbol{y} \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ *is said to be invariant under the group of transformations* $G := \{g : \mathbb{R}^n \to \mathbb{R}^n\}$ *if, for every* $g \in \mathfrak{G}$ *and* $\boldsymbol{\theta} \in \Theta$*, there exists a unique* $\boldsymbol{\theta}^* \in \Theta$ *such that* $\boldsymbol{X} = g(\boldsymbol{Y})$ *has density* $f(\boldsymbol{x} \mid \boldsymbol{\theta}^*) \in \mathfrak{F}$*. In such a situation,* $\boldsymbol{\theta}^*$ *will be denoted* $\bar{g}(\boldsymbol{\theta})$*.*

There are two consequences of applying invariance here. The first is a new criterion:

Criterion 7 - Group invariance: *If all models are invariant under a group of transformations* G_0*, then the conditional distributions,* $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})$*, should be chosen in such a way that the conditional marginal distributions*

$$f_i(\boldsymbol{y} \mid \boldsymbol{\alpha}) = \int f_i(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_i)\, \pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})\, d\boldsymbol{\beta}_i, \tag{5}$$

are also invariant under G_0. (Here, $(\boldsymbol{\alpha}, \boldsymbol{\beta}_i, i)$ would correspond to $\boldsymbol{\theta}$ in the definition of invariance.)

Indeed, the $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})$ could hardly be called objective model selection priors if they eliminated an invariance structure that was possessed by all of the original models. This can also be viewed as a formalization of the Jeffreys (1961) requirement that the prior for a non-null parameter should be "centered at the simple model."

The second use of invariance is in determining the objective prior for the common model parameters $\pi_i(\boldsymbol{\alpha})$. Since all of the marginal models, $f_i(\boldsymbol{y} \mid \boldsymbol{\alpha})$, will be invariant under G_0 if the Group invariance criterion is applied, there are compelling reasons to choose the prior

$$\pi_i(\boldsymbol{\alpha}) = \pi^H(\boldsymbol{\alpha}) \qquad \text{for all } i\,, \tag{6}$$

where $\pi^H(\cdot)$ is the right-Haar density corresponding to the group G_0. The reason is given in Berger, Pericchi and Varshavsky (1998), namely that under commonly satisfied conditions (satisfied for the variable selection problem – see Result 2 in Section 3), use of a common $\pi^H(\boldsymbol{\alpha})$ for all marginal models then ensures exact predictive matching among the models for the minimal training sample size, as in the example given in Section 2.4.

The most surprising feature of this result is that $\pi^H(\boldsymbol{\alpha})$ is typically improper (and hence could be multiplied by an arbitrary constant) and yet, if the same $\pi^H(\boldsymbol{\alpha})$ is used for all marginal models, the prior is appropriately calibrated across models in the strong sense of exact predictive matching. (For any improper prior that occurred in both the numerator and denominator of a Bayes factor, any arbitrary multiplicative constant would obviously cancel but, this is not nearly as compelling a justification as exact predictive matching.) The right-Haar prior is also the objective estimation prior for such models, and so has been extensively studied in invariant situations.

Thus, for invariant models, the combination of the Group invariance criterion and (exact) Predictive matching criterion allows complete specification of the prior for $\boldsymbol{\alpha}$ in all models. It is also surprising that this argument does not require orthogonality of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_i$ (i.e., cross-information of zero in the Fisher information matrix) which, since Jeffreys (1961), has been viewed

as a necessary condition to say that one can use a common prior for $\boldsymbol{\alpha}$ in different models (see, e.g., Hsiao, 1997; Kass and Vaidyanathan, 1992).

There might be concern here as to use of improper priors, even if they are exact predictive matching, especially because of the discussion in Section 2.2. This concern is obviated by the realization that use of any series of proper priors approximating $\pi^H(\alpha)$ will, in the limit, yield Bayes factors equal to that obtained directly from $\pi^H(\alpha)$; see Lemma 1 in Appendix 1.

3. Objective prior distributions for variable selection in normal linear models.

3.1. *Introduction.* We now turn to a particular scenario – variable selection in normal linear models – to illustrate application of the criterion in Section 2. Consider a response variable Y known to be explained by k_0 variables (e.g. an intercept) and by some subset of p other possible explanatory variables. This can formally be stated as a model selection problem with the following 2^p competing models for data $\boldsymbol{y} = (y_1, \ldots, y_n)$:

$$
\begin{aligned}
&M_0 : f_0(\boldsymbol{y} \mid \boldsymbol{\beta}_0, \sigma) = \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0, \sigma^2\boldsymbol{I}) \\
&M_i : f_i(\boldsymbol{y} \mid \boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}_i, \sigma^2\boldsymbol{I}),\ i = 1, \ldots, 2^p - 1\,,
\end{aligned} \tag{7}
$$

where $\boldsymbol{\beta}_0$, σ, and the $\boldsymbol{\beta}_i$ are unknown. Here $\boldsymbol{X}_0$ is a $n \times k_0$ design matrix corresponding to the k_0 variables common to all models; often $\boldsymbol{X}_0 = \mathbf{1}$ so M_0 contains only the intercept. Finally, the $\boldsymbol{X}_i$ are $n \times k_i$ design matrices corresponding to k_i of the p other possible explanatory variables. We make the usual assumption that all design matrices are full rank (without loss of generality). Note that, if the covariance matrix is of the form $\sigma^2\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda}$ known, simply transform $\boldsymbol{Y}$ so that the covariance matrix is proportional to the identity; note that this does not alter the meaning of the $\boldsymbol{\beta}$'s and hence the meaning of the models. Also, setting $\boldsymbol{\alpha} = (\boldsymbol{\beta}_0, \sigma)$ and $N = 2^p$ puts this in the general framework discussed earlier, with M_0 being the null model.

The primary development is for the most common situation of σ unknown and $k_0 \geq 1$, but the simpler cases where either σ is known or $k_0 = 0$ (i.e., the null model only contains the error term) are briefly treated in Section 3.5.

In this setting and following Jeffreys desiderata, Zellner and Siow (1980) recommended use of common objective estimation priors for $\boldsymbol{\alpha}$ (after orthogonalization) and multivariate Cauchy priors for $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\alpha})$, centered at zero and with prior scale matrix $\sigma^2 n(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}$; a similar scale matrix was used in Zellner (1986) for the g-prior.

3.2. *Proposed prior (the 'robust prior').* It is useful to first write down the specific form of the prior that will result from applying the criteria. Indeed, under model M_i, the prior is of the form

$$\begin{aligned} \pi_i^R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma) &= \pi(\boldsymbol{\beta}_0, \sigma) \times \pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma) \\ &= \sigma^{-1} \times \int_0^\infty \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, g\,\boldsymbol{\Sigma}_i)\, p_i^R(g)\, dg, \end{aligned} \tag{8}$$

where $\boldsymbol{\Sigma}_i = Cov(\hat{\boldsymbol{\beta}}_i) = \sigma^2 (\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$ is the covariance of the maximum likelihood estimator of $\boldsymbol{\beta}_i$, with

$$\boldsymbol{V}_i = (\boldsymbol{I}_n - \boldsymbol{X}_0(\boldsymbol{X}_0^t\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^t)\boldsymbol{X}_i \tag{9}$$

and

$$p_i^R(g) = a\,[\rho_i(b+n)]^a\,(g+b)^{-(a+1)} 1_{\{g > \rho_i(b+n)-b\}}\,, \tag{10}$$

$$\text{with} \quad a > 0, \quad b > 0, \quad \text{and} \quad \rho_i \geq \frac{b}{b+n}\,. \tag{11}$$

Note that these conditions ensure that $p_i^R(g)$ is a proper density and g is positive (necessary in (8)), so that $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ is proper, satisfying the first part of the Basic criterion of Section 2.2. The particular choices of hyperparameters that we favor are discussed in Section 3.4.

The prior (8) has its origins in the *robust prior* introduced by Strawderman (1971) and Berger (1980, 1985), for estimating a k-variate normal mean $\boldsymbol{\beta}$ in the sampling scheme $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_k(\boldsymbol{\beta}, \boldsymbol{\Sigma})$. More precisely, the full conditional of $\boldsymbol{\beta}_i$ induced by (8) generalizes the above mentioned *robust prior* considering the sampling distribution of the maximum likelihood estimator, namely $\hat{\boldsymbol{\beta}}_i \sim \mathcal{N}_{k_i}(\boldsymbol{\beta}_i, \sigma^2 (\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1})$. The primary reasons for Strawderman (1971) and Berger (1980, 1985) to consider such priors was that it results in closed

form inferences, including closed form Bayes factors, and results in estimates that are robust in various senses. For this reason, we continue the tradition of calling (8) the *robust prior* and use a superindex R to denote it. Note also that priors of this form have been previously considered. The priors proposed by Liang et al. (2008) are particular cases with $a = 1/2$, $b = 1$, $\rho_i = 1/(1+n)$ (the hyper-g prior) and $a = 1/2$, $b = n$, $\rho_i = 1/2$ (the hyper-g/n prior). The prior in Cui and George (2008) has $a = 1$, $b = 1$, $\rho_i = 1/(1 + n)$. The original Berger's prior for robust estimation is the particular case with $a = 1/2$, $b = 1$, $\rho_i = (k_i + 1)/(k_i + 3)$; closely related priors are those of Maruyama and Strawderman (2010); Maruyama and George (2008).

Finally, it is useful to note that $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ behaves in the tails as a multivariate Student distribution (already noticed for a particular case in Berger, 1980, and the reason for its robust estimation properties).

PROPOSITION 1. *Writing* $\|\boldsymbol{\beta}_i\|^2 = \boldsymbol{\beta}_i^t(\boldsymbol{V}_i^t\boldsymbol{V}_i)\boldsymbol{\beta}_i$,

$$\lim_{\|\beta_i\|^2\to\infty} \frac{\pi_i^R(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \sigma)}{St_{k_i}(\boldsymbol{\beta} \mid \mathbf{0},\, (a\,\Gamma(a))^{1/a}\,\rho_i\,\boldsymbol{B}^*(b,\sigma)/a,\, 2a)} = 1,$$

where $\boldsymbol{B}^*(b,\sigma) = \sigma^2(b+n)(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$.

PROOF. See Appendix 2. □

In the model selection scenario, the thickness of the prior tails is related to the information consistency criteria, and is the reason Jeffreys (1961) used a Cauchy as the prior for testing a normal mean. Also, using this result, we can see that $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ has close connections with the Zellner-Siow priors; in fact, for $a = 1/2$, $b = n$, $\rho_i = 2/\pi$ and large n, $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ and the Zellner-Siow priors have exactly the same tails.

3.3. *Justification of model selection priors of the form (8).* We will use the Group invariance criterion and Predictive matching criterion (along with practical computational considerations) to justify use of model selection priors of the form (8). We first justify the use of $\pi^R(\boldsymbol{\beta}_0, \sigma) = 1/\sigma$ for the common parameters and then justify the choice $\pi_i^R(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \sigma)$ for the model specific parameters.

3.3.1. *Justification of the prior for the common parameters.* It is convenient, in this section, to consider a more general class of conditional priors,

$$\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma) = \sigma^{-k_i} h_i(\frac{\boldsymbol{\beta}_i}{\sigma})\,, \tag{12}$$

where h_i is any proper density with support $\mathcal{R}^{k_i}$. The robust prior is the particular case

$$h_i^R(\boldsymbol{u}) = \int \mathcal{N}_{k_i}(\boldsymbol{u} \mid \mathbf{0},\, g\,(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1})\, p_i^R(g)\, dg\,. \tag{13}$$

It is shown, in Appendix 3, that all models in (7) are invariant under the group of transformations

$$G_0 = \{g = (c, \boldsymbol{b}) \in (0, \infty) \times \mathcal{R}^{k_0} : g(\boldsymbol{y}) \to c\boldsymbol{y} + \boldsymbol{X}_0\boldsymbol{b}\}.$$

The following establishes a necessary and sufficient condition on the conditional prior $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ for the Group invariance criterion to hold for this group.

Result 1. *The conditional marginals*

$$f_i(\boldsymbol{y} \mid \boldsymbol{\beta}_0,\, \sigma) = \int \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_i\boldsymbol{\beta}_i,\, \sigma^2\boldsymbol{I})\, \pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma)\, d\boldsymbol{\beta}_i \tag{14}$$

are invariant under G_0 if and only if $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0,\, \sigma)$ has the form (12).

Proof. See Appendix 3. □

Based on the Group invariance criterion, Result 1 implies that, conditionally on the common parameters $\boldsymbol{\beta}_0$ and σ, $\boldsymbol{\beta}_i$ must be scaled by σ, centered at zero and not depend on $\boldsymbol{\beta}_0$ (as was argued for simple normal testing in Jeffreys, 1961). Note, in particular, that the robust prior in (8) satisfies the Group invariance criterion (although it is not the only prior that does so).

Next, since each marginal model $f_i(\boldsymbol{y} \mid \boldsymbol{\beta}_0,\, \sigma)$ resulting from a prior in (12) is invariant with respect to G_0, the suggestion from Berger, Pericchi and Varshavsky (1998) is to use the right-Haar density for the common parameters $(\boldsymbol{\beta}_0, \sigma)$, namely

$$\pi_i(\boldsymbol{\beta}_0, \sigma) = \pi^H(\boldsymbol{\beta}_0, \sigma) = \sigma^{-1},$$

the right-Haar prior for the location-scale group. Using this, the overall model prior would be of the form

$$\pi_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma) = \sigma^{-1-k_i}\, h_i(\frac{\boldsymbol{\beta}_i}{\sigma}). \tag{15}$$

The justification for the right-Haar prior in Berger, Pericchi and Varshavsky (1998) depends, however, on showing that it is predictive matching, in the sense described in the following result.

RESULT 2. *For M_i, let the prior $\pi_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma)$ be of the form* (15)*, where h_i is symmetric about zero. Then all model/prior pairs $\{M_i, \pi_i\}$ are exact predictive matching for $n^* = k_0 + 1$.*

PROOF. See Appendix 4. □

The conclusion of the above development is that the Group invariance criterion and Predictive matching criterion imply that model selection priors should be of the form (15), with h_i symmetric about zero. It would thus appear that the robust prior satisfies these criteria, as (13) is clearly symmetric about zero. (Any scale mixture of Normals would also satisfy these criteria, since the resulting $h(\cdot)$ would be symmetric about 0.) Note, however, that h_i^R has scale matrix proportional to $(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$, and $\boldsymbol{V}_i$ in (9) requires both $\boldsymbol{X}_0$ and $\boldsymbol{X}_i$, which would seem to indicate that a sample size of $k_0 + k_i$ is required. Hence, Result 2 would seem to apply to the robust prior only if $k_i = 1$.

This is a situation, however, where the definition of a minimal sample size is somewhat ambiguous. For instance, suppose one were presented $\boldsymbol{X}_0$ and $\boldsymbol{X}_i$ for $k_0 + k_i$ observations for each model M_i, but that only $k_0 + 1$ of the y_i were reported for all models, with the rest being missing data. This is still a minimal sample size in the sense that it is the smallest collection of y_i for which all marginal densities exist for the robust prior, and now Result 2 applies to say that the robust prior is predictive matching for all models.

3.3.2. *Justification of the prior for the model specific parameters.* While the robust prior is thus validated as satisfying the Group invariance criterion and a version of the Predictive matching criterion, there are many other

model selection priors of the form (15) which also satisfy these criteria. There are additional reasons, however, to focus on the robust priors with $h_i^R(\boldsymbol{u})$ of the form (13). The first is that only scale mixtures of normals seem to have any possibility of yielding Bayes factors that have closed form. While we have not focused on this as a necessary criterion, it is an attractive enough property to justify the restriction. There are, however, two other features of (13) that need justification: the use of the mixture density $p_i^R(g)$, and the choice of the conditional scale matrix $(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$.

The mixture density $p_i^R(g)$ encompasses virtually all of the mixtures that have been found which can lead to closed form expressions for Bayes factors; for example, Zellner-Siow priors are scale mixtures of normals but with a different mixing density which does not lead to close-form expressions. (The choice of mixing density in Maruyama and George, 2008, is a very interesting exception, in that it leads to a closed form expression for a different reason than does $p_i^R(g)$.) So, while not completely definitive, $p_i^R(g)$ is an attractive choice. The choice of $(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$ as the conditional scale matrix seems much more arbitrary, but there is one standard argument and one surprising argument in its favor.

The standard argument is the Measurement Invariance criterion; if the conditional scale matrix is chosen to be $(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$, it is easy to see that Bayes factors will be unaffected by changes in the units of measurement of either $\boldsymbol{y}$ or the model parameters. But there are many other choices of the conditional scale matrix which also have this property.

A quite surprising predictive matching result that supports use of $(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$ as the conditional scale matrix is as follows.

RESULT 3. *For M_i, let the prior be as in* (15) *where h_i is the scale mixture of normals in* (13)*. The priors are then null predictive matching and dimensional predictive matching for samples of size $k_0 + k_i$, and no choice of the conditional scale matrix other than $(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$ (or a multiple) can achieve this predictive matching.*

PROOF. See Appendix 5. □

This is surprising, in that it is a predictive matching result for larger sam-

ple sizes (k_0+k_i) than are encountered in typical predictive matching results, such as Result 2. That it only holds for conditional scale matrices proportional to $(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$ is also surprising, but does strongly support choosing a prior of the form (8).

3.4. *Choosing the hyperparameters for $p_i^R(g)$.*

3.4.1. *Introduction.* The Bayes factor of M_i to M_0 arising from the robust prior π_i^R in (8) can be compactly expressed as the following function of the hyperparameters a, b and ρ_i:

$$B_{i0} = Q_{i0}^{-\frac{n-k_0}{2}} \frac{2a}{k_i+2a} \left[\rho_i(n+b)\right]^{-k_i/2} \mathrm{AP}_i, \tag{16}$$

where AP_i is the hypergeometric function of two variables (see Weisstein, 2009), or Apell hypergeometric function

$$\mathrm{AP}_i = \mathrm{F}_1\Big[a+\frac{k_i}{2};\ \frac{k_i+k_0-n}{2},\ \frac{n-k_0}{2};\ a+1+\frac{k_i}{2};\ \frac{(b-1)}{\rho_i(b+n)};\ \frac{b-Q_{i0}^{-1}}{\rho_i(b+n)}\Big],$$

and $Q_{i0} = SSE_i/SSE_0$ is the ratio of the sum of squared errors of models M_i and M_0. The details of this computation are given in Appendix 6.

Having a closed form expression for Bayes factors is not one of our formal criteria for model selection priors, but it is certainly a desirable property, especially when realizing that one is dealing with 2^p models in variable selection.

The values for the hyperparameters that will be recommended are $a = 1/2$, $b = 1$ and $\rho_i = (k_i + k_0)^{-1}$. The arguments justifying this specific recommendation follow.

3.4.2. *Implications of the consistency criteria.* The consistency criteria of Section 2.1 provide considerable guidance as to the choice of a, b and the ρ_i. In particular, they lead to the following result.

Result 4. *The three consistency criterion of Section 2.3 are satisfied by the robust prior if a and ρ_i do not depend on n, $\lim_{n\to\infty}\frac{b}{n} = c \geq 0$, $\lim_{n\to\infty}\rho_i\,(b+n) = \infty$, and $n \geq k_i + k_0 + 2a$.*

This result follows from (18), (20), and (22) below, which are presented as separate results because they can be established in more generality than simply for the robust prior.

Use of model selection consistency. Suppose M_i is the true model, and consider any other model M_j. A key assumption for model selection consistency (Fernández, Ley and Steel, 2001) is that, asymptotically, the design matrices are such that the models are differentiated, in the sense that

$$\lim_{n\to\infty} \frac{\boldsymbol{\beta}_i^t \boldsymbol{V}_i^t (\boldsymbol{I} - \boldsymbol{P}_j) \boldsymbol{V}_i \boldsymbol{\beta}_i}{n} = b_j \in (0, \infty)\,, \tag{17}$$

where $\boldsymbol{P}_j = \boldsymbol{V}_j (\boldsymbol{V}_j^t \boldsymbol{V}_j)^{-1} \boldsymbol{V}_j^t$.

RESULT 5. *Suppose (17) is satisfied and that the priors* $\pi_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma)$ *are of the form* (15)*, with* $h_i(\boldsymbol{u}) = \int \mathcal{N}_{k_i}(\boldsymbol{u} \mid \boldsymbol{0},\, g\,(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1})\, p_i(g)\, dg$*. If the* $p_i(g)$ *are proper densities such that*

$$\lim_{n\to\infty} \int_0^\infty (1+g)^{-k_i/2}\, p_i(g)\, dg = 0\,,$$

model selection consistency will result.

PROOF. The proof follows directly from the proof of Theorem 3 in Liang et al. (2008) and is, hence, omitted. □

COROLLARY 1. *The prior distributions in* (8) *are model selection consistent if*

$$\lim_{n\to\infty} \rho_i\, (b+n) = \infty\,. \tag{18}$$

PROOF. See Appendix 7. □

Use of intrinsic prior consistency. Related to (17) is the condition that

$$\lim_{n\to\infty} \frac{1}{n}\, \boldsymbol{V}_l^t \boldsymbol{V}_l = \boldsymbol{\Xi}_l\,, \tag{19}$$

for some positive definite matrix $\boldsymbol{\Xi}_l$. This would trivially happen if either there is a fixed design with replicates, or when the covariates arise randomly from a fixed distribution having second moments.

RESULT 6. *If (19) holds,*

$$a \text{ and } \rho_i \text{ do not depend on } n, \quad \text{and} \quad \frac{b}{n} \to c\,, \tag{20}$$

then the conditional robust prior $\pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ *in (8) converges to the fixed intrinsic prior*

$$\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma) = \int_0^\infty \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, g^*\,\sigma^2 \boldsymbol{\Xi}^{-1})\, p_i(g^*)\, dg^*, \tag{21}$$

where $p_i(g^*) = a[\rho_i(c+1)]^a\,(g^*+c)^{-(a+1)} 1_{\{g^* > \rho_i(c+1)-c\}}$.

PROOF. Changing variables to $g^* = g/n$, the integral in (8) becomes

$$\int_0^\infty \mathcal{N}_{k_i}\left(\boldsymbol{\beta}_i \mid \mathbf{0}, g^*\sigma^2 \left(\frac{1}{n}\boldsymbol{V}_l^t \boldsymbol{V}_l\right)^{-1}\right) a \left[\rho_i\left(\frac{b}{n}+1\right)\right]^a \times \left(g^* + \frac{b}{n}\right)^{-(a+1)} 1_{\{g^* > \rho_i(\frac{b}{n}+1)-\frac{b}{n}\}} dg^*\,.$$

For large n and using (19) and (20), it is easy to find an integrable function dominating the integrand, so the dominated convergence theorem can be applied to interchange the integral and limit, yielding the result. □

Use of information consistency. For the variable selection problem, it is easy to see that

$$\sup_{\boldsymbol{\beta}_l, \boldsymbol{\beta}_0, \sigma} f_l(\boldsymbol{y} \mid \boldsymbol{\beta}_0,\, \boldsymbol{\beta}_l,\, \sigma) = (2\pi \mathrm{SSE}_l/n)^{-n/2} \exp\left(-n/2\right)$$

for model M_l. Hence, for any given data set $\boldsymbol{y}$ the estimated likelihood ratio in (4) is

$$\Lambda_{i0}(\boldsymbol{y}) = Q_{i0}(\boldsymbol{y})^{-n/2}\,,$$

where $Q_{i0}(\boldsymbol{y})$ is the ratio of the residual sum of squares of the two models for $\boldsymbol{y}$. Therefore, having a sequence of data vectors $\{\boldsymbol{y}_m\}$ such that $\lim_{m\to\infty} \Lambda_{i0}(\boldsymbol{y}_m) = \infty$ is equivalent to having a sequence of data vectors such that $\lim_{m\to\infty} Q_{i0}(\boldsymbol{y}_m) \to 0$.

RESULT 7. *If* $\rho_i \geq b/(b+n)$, *the prior in* (8) *results in an information consistent Bayes factor for* M_i *versus* M_0, *if and only if*

$$n \geq k_i + k_0 + 2a\,. \tag{22}$$

PROOF. See Appendix 8. □

3.4.3. *Specific choices of hyperparameters.*

The choice of a. Note that, with $k_i > k_j$ and $n \geq k_i + k_0 + 1$, the Bayes factor B_{ij} between M_i and M_j exists. It is desirable to have information consistency for all such sample sizes, in which case (22) would require $a \leq 1/2$. The choice $a = 1/2$ is attractive, in that it coincides with the choice in Berger (1985) and, with this choice, π_l^R has Cauchy tails, as do the popular proposals of Jeffreys (1961) and Zellner and Siow (1980, 1984).

Additional motivation for this choice can be found by studying the behavior of B_{i0} when the information favors M_0, in the sense that $Q_{i0} \to 1$. Indeed, Forte (2011) shows that the limiting value of B_{i0} is then bounded above by $2a/(2a + k_i)$ for any sample size, including a small sample size such as $k_0 + k_i + 1$. A small value of a would imply strong evidence in favor of M_0, which does not seem reasonable when the sample size is small. In contrast, the recommended choice would yield a bound of $1/(1 + k_i)$, which certainly favors M_0, but in a sensibly modest fashion when the sample size is small.

The choice of b. To understand the effect of b and the ρ_i on the robust prior, it is useful to begin by considering the approximating intrinsic prior in Result 6, which depends on the hyperparameters only through the mixing distribution $p_i^R(g^*)$, which for $a = 1/2$ is given by (when $b/n \to c$)

$$p_i^R(g^*) = \frac{1}{2}[\rho_i(c+1)]^{1/2}\,(g^* + c)^{-3/2} 1_{\{g^* > \rho_i(c+1) - c\}}\,. \tag{23}$$

This is a very flat tailed distribution with median $4\rho_i(1 + c) - c$. Because it is so flat-tailed, the choice of c in $(g^* + c)^{-3/2}$ is not particularly influential, so that the main issue is the choice of the median. For selecting a median, however, ρ_i and c are confounded – i.e., we do not need both. For simplicity, therefore, we will choose $c = 0$ (i.e., b such that $b/n \to 0$).

If $b/n \to c = 0$, the intrinsic prior does not depend at all on b. Furthermore, there is very little dependence on b, in this case, for the actual robust prior, as was verified for moderate and small n in Forte (2011) through an extensive numerical study.

Since any choice of b for which $b/n \to 0$ makes little difference, it would be reasonable to make such a choice based on pragmatic considerations. In this

regard, note that the choice $b = 1$ has a notable computational advantage, in that the hypergeometric function of two variables, AP_i, then becomes the standard hypergeometric function of one variable (Abramowitz and Stegun, 1964). We thus choose $b = 1$.

The choice of ρ_i. This is the most difficult choice to make, since there is only limited guidance from the various criteria. To review (and assuming $b = 1$), we have that $\rho_i \geq 1/(1+n)$ (so that $g > 0$); $\lim_{n\to\infty} \rho_i(1+n) = \infty$ (for model selection consistency); and ρ_i should not depend on n (for there to be a limiting intrinsic prior). Also note that n is necessarily greater than or equal to $k_0 + k_i$ for the robust prior and marginal likelihood to exist; supposing we wish to choose ρ_i so that the conditions are satisfied for all such n, these restrictions only imply that

$$\rho_i \text{ must be a constant (independent of } n) \text{ and } \rho_i \geq 1/(1+k_0+k_i)\,.$$

We present two arguments below for the specific choice $\rho_i = 1/(k_0 + k_i)$.

Argument 1. Consider the Bayes factor B_{i0} of M_i to M_0. In Result 3, it was established that $B_{i0} = 1$ for a sample of size $n = k_i + k_0$, but a natural question is – what should we expect for a sample of size $n = k_i + k_0 + 1$? Can a single additional observation provide much information to discriminate between M_i and M_0? Intuition says no. To quantify the intuition, consider the situation in which $Q_{i0} \to 1$, which corresponds to information being as supportive as possible of M_0. It is straightforward to show that, when $n = k_i + k_0 + 1$,

$$\lim_{Q_{i0}\to 1} B^R_{i0} = \frac{1}{k_i+1}[\rho_i(k_i+k_0+2)]^{-k_i/2}\,. \tag{24}$$

As we should not expect a single extra observation to provide very strong evidence, even in the case that $Q_{i0} \to 1$, the implication is that we should choose ρ_i to be as small as is reasonable. The choice $\rho_i = 1/(k_0 + k_i + 1)$ is the minimum value of ρ_i and is, hence, certainly a candidate.

Argument 2. Consider the intrinsic prior defined by (21) and (23). Note that we have chosen $c = 0$ (through the choice of $b = 1$) and, after making the transformation $\tilde{g} = g^*/\rho_i$, the intrinsic prior can be written

$$\pi_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma) = \sigma^{-1} \times \int_0^\infty \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, \tilde{g}\,\rho_i\,\boldsymbol{\Xi}^{-1})\, p_i(\tilde{g})\, d\tilde{g}\,, \tag{25}$$

where $p_i(\tilde{g}) = (1/2)(\tilde{g})^{-3/2}1_{\{\tilde{g}>1\}}$. Thus we see that, in the intrinsic prior approximation to the robust prior, ρ_i can be interpreted as simply a scale factor to the conditional covariance matrix. This helps, in that there have been previous suggestions related to 'unit information priors' (Kass and Wasserman, 1995; Berger, Bayarri and Pericchi, 2012). For instance, Berger, Bayarri and Pericchi (2012) considers the group means problem defined as follows: the observations are

$$y_{ij} = \mu_i + \epsilon_{ij}\,, \quad i = 1, \ldots, k \quad \text{and} \quad j = 1, \ldots, r\,,$$

with i.i.d. $\epsilon_{ij} \sim N(\cdot \mid 0, \sigma^2)$. Thus there are k different means, μ_i, and r replicate observations for each. Applying the robust prior to this example (considering the full model with all μ_i) results in a conditional covariance matrix in (25) of $\rho\, k\, \boldsymbol{I}$, which is much too diffuse if k is large and ρ is not small. Selecting $\rho = 1/k$, on the other hand, restores a 'unit information' prior. Here k_0=0, so the choice $\rho = 1/k$ is equivalent to the overall choice $\rho_i = 1/(k_0 + k_i)$. This overall choice is obviously very close to earlier suggested $1/(k_0 + k_i + 1)$.

3.5. *Two simpler cases.* We conclude with discussion of the modifications of the robust prior that are needed when $\boldsymbol{\beta}_0 = 0$ or when σ is known.

3.5.1. *When $\boldsymbol{\beta}_0 = \mathbf{0}$ and σ is unknown.* When $\boldsymbol{\beta}_0 = \mathbf{0}$, the robust prior distribution is

$$\pi_i^R(\boldsymbol{\beta}_i, \sigma) = \pi(\sigma) \times \pi_i^R(\boldsymbol{\beta}_i \mid \sigma) = \sigma^{-1} \times \int_0^\infty \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, g\, \boldsymbol{\Sigma}_i)\, p_i^R(g)\, dg,$$

where $\boldsymbol{\Sigma}_i = Cov(\hat{\boldsymbol{\beta}}_i) = \sigma^2\, (\boldsymbol{X}_i^t \boldsymbol{X}_i)^{-1}$, the covariance of the maximum likelihood estimator of $\boldsymbol{\beta}_i$ and, as before,

$$p_i^R(g) = a[\rho_i(b+n)]^a\, (g+b)^{-(a+1)}, \quad g > \rho_i(b+n) - b.$$

The corresponding Bayes factor is as in (16) with $k_0 = 0$; when we choose $a = 1/2$, $b = 1$ and $\rho_i = 1/(k_i + k_0)$ it assumes the simpler form in (26), again with $k_0 = 0$.

In regards to the Group invariance criterion, when $\boldsymbol{\beta}_0 = 0$ the models are invariant under the scale group of transformations, $G_0 = \{\boldsymbol{y} \to c\boldsymbol{y},\ \ c > 0\}$,

and it is easy to show that $\pi(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ still needs to be a scale prior, as in (12), to preserve the invariance structure; also, the use of $\pi(\sigma) = 1/\sigma$ is again justified by predictive matching, as it is the Haar prior for the group. Null and dimensional predictive matching also hold as well as the various consistency criteria.

3.5.2. *When σ is known and $\boldsymbol{\beta}_0 \neq 0$.* When σ is known, the robust prior becomes

$$\pi_i^R(\boldsymbol{\beta}_i, \boldsymbol{\beta}_0, \sigma) = \pi(\boldsymbol{\beta}_0) \times \pi_i^R(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0) \propto \int_0^\infty \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, g\,\boldsymbol{\Sigma}_i)\, p_i^R(g)\, dg,$$

where $\boldsymbol{\Sigma}_i = Cov(\hat{\boldsymbol{\beta}}_i) = \sigma^2\,(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$, and $p_i^R(g)$ is as before.

The models are now invariant under the location group $G_0 = \{\boldsymbol{y} \to \boldsymbol{y} + \boldsymbol{X}_0\,\boldsymbol{b},\ \ \boldsymbol{b} \in \mathcal{R}^{k_0}\}$, and it is easy to show that $\pi(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0)$ just needs to be independent of $\boldsymbol{\beta}_0$ to preserve the invariance structure; the use of the Haar prior $\pi(\boldsymbol{\beta}_0) = 1$ is again justified through predictive matching arguments.

The Bayes factor can be expressed as

$$B_{i0} = \int_0^\infty (g+1)^{-k_i/2} \Lambda_{0i}^{\left(\frac{1}{g+1}-1\right)} p_i(g) dg\,,$$

where $\Lambda_{0i} = \exp\left(-[SSE_0 - SSE_i]/(2\sigma^2)\right)$. This is curiously difficult to express in closed-form in general but, for our preferred choice $b = 1$, change of variables to $h = 1/(1+g)$ yields

$$\begin{aligned} B_{i0} &= \int_0^\infty (g+1)^{-k_i/2} \Lambda_{0i}^{\left(\frac{1}{g+1}-1\right)} a(\rho_i(1+n))^a\,(g+1)^{-(a+1)} 1_{\{g > \rho_i(1+n)-1\}} dg \\ &= a(\rho_i(1+n))^a \Lambda_{0i}^{-1} \int_0^{1/[\rho_i(1+n)]} h^{(a-1+k_i/2)} e^{-h[SSE_0 - SSE_i]/(2\sigma^2)} dh \\ &= a(\rho_i(1+n))^a \Lambda_{0i}^{-1} \left(\frac{[SSE_0 - SSE_i]}{2\sigma^2}\right)^{-(a-2+\frac{k_i}{2})} \\ &\qquad \times \left(\Gamma\left[a + \frac{k_i}{2}\right] - \Gamma\left[a + \frac{k_i}{2}, \frac{[SSE_0 - SSE_i]}{2\sigma^2 \rho_i (1+n)}\right]\right), \end{aligned}$$

where $\Gamma(\nu_1, \nu_2)$ is the incomplete gamma function,

$$\Gamma(\nu_1, \nu_2) = \int_{\nu_2}^\infty t^{\nu_1 - 1} e^{-t}\, dt\,.$$

All of the properties of the procedures for the σ unknown case also hold here, except for null predictive matching.

4. Methodological summary for variable selection. Although the primary purpose of the paper was to develop the criteria for choice of model selection priors and study their implementation in an example, the methodological results obtained for the problem of variable selection in the normal linear model, as outlined in Section 3.1, are of interest in their own right. For ease of use, we summarize these developments here.

Using the notation of Section 3.1, the prior distribution recommended for the parameters under model M_i is

$$\pi_i^R(\boldsymbol{\beta}_0, \boldsymbol{\beta}_i, \sigma) = \sigma^{-1} \times \int_0^\infty \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid \mathbf{0}, g\,\boldsymbol{\Sigma}_i)\, p_i^R(g)\, dg,$$

where $\boldsymbol{\Sigma}_i = \sigma^2\,(\boldsymbol{V}_i^t \boldsymbol{V}_i)^{-1}$, $\boldsymbol{V}_i = (\boldsymbol{I}_n - \boldsymbol{X}_0(\boldsymbol{X}_0^t\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^t)\boldsymbol{X}_i$, and

$$p_i^R(g) = \frac{1}{2}\,\left[\frac{(1+n)}{(k_i+k_0)}\right]^{\frac{1}{2}}\,(g+1)^{-3/2} 1_{\{g>(k_i+k_0)^{-1}(1+n)-1\}}\,.$$

The resulting Bayes factors have closed form expressions in terms of the the hypergeometric function, namely

(26)

$$B_{i0} = \left[\frac{n+1}{k_i+k_0}\right]^{-\frac{k_i}{2}} \frac{Q_{i0}^{-\frac{n-k_0}{2}}}{k_i+1}\,{}_2F_1\Big[\frac{k_i+1}{2}; \frac{n-k_0}{2}; \frac{k_i+3}{2}; \frac{(1-Q_{i0}^{-1})(k_i+k_0)}{(1+n)}\Big],$$

where ${}_2F_1$ is the standard hypergeometric function (see Abramowitz and Stegun, 1964) and $Q_{i0} = SSE_i/SSE_0$ is the ratio of the sum of squared errors of models M_i and M_0.

To implement Bayesian model selection through (2), one also needs the prior odds ratios P_{j0}. A recommended objective Bayesian choice of these odds ratios for the variable selection problem is $P_{j0} = k_j!(p-k_j)!/p!$. For extensive discussion and earlier references see Scott and Berger (2010).

References.

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover.

Bayarri, M. J. and García-Donato, G. (2008). Generalization of Jeffreys Divergence-Based Priors for Bayesian Hypothesis Testing. *Journal of the Royal Statistical Society: Series B* **70** 981–1003.

Berger, J. O. (1980). A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean. *The Annals of Statistics* **8** 716–761.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer.

Berger, J. O., Bayarri, M. J. and Pericchi, L. R. (2012). The Effective Sample Size. *Econometric Reviews* (submitted).

Berger, J. O., Ghosh, J. K. and Mukhopadhyay, N. (2003). Approximations and Consistency of Bayes Factors as Model Dimension Grows. *Journal of Statistical Planning and Inference* **112** 241 - 258.

Berger, J. O. and Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association* **91** 109-122.

Berger, J. O., Pericchi, L. R. and Varshavsky, J. A. (1998). Bayes Factors and Marginal Distributions in Invariant Situations. *Sankhya: The Indian Journal of Statistics, Series A* **60** 307–321.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison. *Lecture Notes-Monograph Series* **38** 135–207.

Casella, G., Girón, F. J., Martínez, M. L. and Moreno, E. (2009). Consistency of Bayesian Procedures for Variable Selection. *The Annals of Statistics* **37** 1207-1228.

Cui, W. and George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference* **138** 888-900.

De Santis, F. and Spezzaferri, F. (1999). Methods for default and roubst Bayesian model Comparison: The Fractional Bayes factor approach. *International Statistical Review* **67** 267-286.

Fernández, C., Ley, E. and Steel, M. F. (2001). Benchmark Priors for Bayesian Model Averaging. *Journal of Political Economics* **100** 381-427.

Forte, A. (2011). Objective Bayesian Criteria for Variable Selection. PhD Thesis, Universidad de Valencia.

Ghosh, J. K. and Samanta, T. (2002). Nonsubjective Bayes Testing: An Overview. *Journal of Statistical Planning and Inference* **103** 205-223.

Guo, R. and Speckman, P. L. (2009). Bayes Factors Consistency in Linear Models. Presented in O'Bayes 09 conference.

Hsiao, C. K. (1997). Approximate Bayes Factors When a Mode Occurs on the Boundary. *Journal of the American Statistical Association* **92** 656–663.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association* **90** 773–795.

Kass, R. E. and Vaidyanathan, S. K. (1992). Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions.

Journal of the Royal Statistical Society. Series B (Methodological) **54** 129–144.

Kass, R. E. and Wasserman, L. (1995). A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association* **90** 928–934.

Laud, P. W. and Ibrahim, J. G. (1995). Predictive Model Selection. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 247–262.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association* **103** 410–423.

Maruyama, Y. and George, E. I. (2008). gBF: A fully Bayes factor with a generalized g-prior. arXiv:0801.4410v2 [stat.ME].

Maruyama, Y. and Strawderman, W. E. (2010). Robust Bayesian variable selection with sub-harmonic priors. arXiv:1009.1926v2 [stat.ME].

Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. *Journal of the American Statistical Association* **93** 1451-1460.

Pérez, J. M. and Berger, J. O. (2002). Expected posterior prior distributions for model selection. *Biometrika* **89** 491-512.

Robert, C. P., Chopin, N. and Rousseau, J. (2009). Harold Jeffreys' theory of probability revisited. *Statistical Science* **24** 141-172.

Scott, J. G. and Berger, J. O. (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *The Annals of Statistics* **38** 2587-2619.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes Factors for Linear and Log-Linear Models with Vague Prior Information. *Journal of the Royal Statistical Society. Series B (Methodological)* **44** 377–387.

Strawderman, W. E. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *The Annals of Mathematical Statistics* **42** 385–388.

Suzuki, Y. (1983). On Bayesian Approach to Model Selection. In *Proceedings of the International Statistical Institute* 288-291. Voorburg, ISI Publications.

Weisstein, E. W. (2009). Appell Hypergeometric Function From Mathworld- A Wolfram web Resource. http://mathworld.wolfram.com/AppellHypergeometricFunction.html.

Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions. In *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti* (A. Zellner, ed.) 389-399. Edward Elgar Publishing Limited.

Zellner, A. and Siow, A. (1980). Posterior Odds Ratio for Selected Regression Hypotheses. In *Bayesian Statistics 1* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585-603. Valencia: Univeristy Press.

Zellner, A. and Siow, A. (1984). *Basic Issues in Econometrics*. Chicago: University of Chicago Press.

APPENDICES

A1. Approximations to improper priors.

LEMMA 1. *Consider* $\pi_i(\boldsymbol{\alpha}) = c_i\psi_i(\boldsymbol{\alpha})$, *where* $\psi_i(\boldsymbol{\alpha})$ *increases monotonically in* i *to* $\pi(\boldsymbol{\alpha})$ *and* $c_i = 1/\int \psi_i(\boldsymbol{\alpha})d\boldsymbol{\alpha} < \infty$. *Then, if* $\int f_l(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\pi(\boldsymbol{\alpha})\,d\boldsymbol{\alpha} < \infty$ *for all densities* $f_l(\boldsymbol{y} \mid \boldsymbol{\alpha})$,

$$\lim_{i\to\infty} \frac{\int f_l(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\pi_i(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}}{\int f_{l'}(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\pi_i(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}} = \frac{\int f_l(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\pi(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}}{\int f_{l'}(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\pi(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}}\,.$$

PROOF.

$$\frac{\int f_l(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\pi_i(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}}{\int f_{l'}(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\pi_i(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}} = \frac{\int f_l(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\psi_i(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}}{\int f_{l'}(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\psi_i(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}} \longrightarrow \frac{\int f_l(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\pi(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}}{\int f_{l'}(\boldsymbol{y} \mid \boldsymbol{\alpha})\,\pi(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}}$$

by the monotone convergence theorem. □

Thus common proper priors can be used to approximate common improper priors and, as the approximation improves, the Bayes factors for the proper priors converge to the Bayes factor for the improper prior; this is why Bayesians have always said that it is not illogical to use an improper prior for a common parameter $\boldsymbol{\alpha}$ in computing a Bayes factor. It is interesting that no conditions are needed in lemma except that the marginal likelihoods exist for the improper prior, which is clearly needed for the Bayes factor to even be defined for the improper prior.

A2. Proof of Proposition 1. This proof requires the following lemma:

LEMMA 2. *If* $m > 1$, $p > 0$, $a > 0$, *and* $k \geq 1$, *then*

$$\lim_{z\to\infty} z^{a+k} \int_0^1 \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z} d\lambda = m^a\,\Gamma(a+k)\,p^{-(a+k)}.$$

PROOF. For $0 < \epsilon < 1$ write

$$\begin{aligned}
&\lim_{z\to\infty} \int_0^1 z^{a+k}\lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\,d\lambda \\
&= \lim_{z\to\infty} \int_0^\epsilon z^{a+k}\lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\,d\lambda \\
&\quad + \lim_{z\to\infty} \int_\epsilon^1 z^{a+k}\lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\,d\lambda\,. \qquad (27)
\end{aligned}$$

Note that

$$\lim_{z\to\infty} z^{a+k}\,\lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z} = 0$$

and the integrand in the last integral in (28) is uniformly bounded over λ and z. It follows from the dominated convergence theorem that the last term is zero, so that

$$\begin{aligned}&\lim_{z\to\infty}\int_0^1 z^{a+k}\lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\,d\lambda\\ &= \lim_{z\to\infty}\int_0^\epsilon z^{a+k}\lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\,d\lambda. \qquad (28)\end{aligned}$$

Next, make the change of variables $t = \lambda/(m-\lambda)$ to get

$$\int_0^\epsilon \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^k e^{-\frac{\lambda}{m-\lambda}\cdot p\cdot z}\,d\lambda = m^a\int_0^{\frac{\epsilon}{m-\epsilon}} \frac{t^{k+a-1}}{(1+t)^{a+1}}\,e^{-t\cdot p\cdot z}\,dt.$$

To bound the integral of interest notice that, for $t\in(0,\epsilon/(m-\epsilon))$,

$$\frac{1}{(1+\epsilon/(m-\epsilon))^{a+1}} \le \frac{1}{(1+t)^{a+1}} \le 1. \qquad (29)$$

By integrating t out from (29) and multiplying the result by $z^{(a+k)}$ we get both an upper and a lower bound for the integral of interest, namely

$$\begin{aligned}&\frac{m^a\,p^{-(a+k)}\left(\Gamma(a+k)-\Gamma(a+k,\,\frac{\epsilon}{m-\epsilon}\,pz)\right)}{(1+\epsilon/(m-\epsilon))^{a+1}}\\ &\le \lim_{z\to\infty} m^a\int_0^{\frac{\epsilon}{m-\epsilon}} z^{a+k}\frac{t^{k+a-1}}{(1+t)^{a+1}}\,e^{-t\cdot p\cdot z}\,dt \qquad (30)\\ &\le m^a\,p^{-(a+k)}\left(\Gamma(a+k)-\Gamma(a+k,\,\frac{\epsilon}{m-\epsilon}\,pz)\right),\end{aligned}$$

where $\Gamma(\nu_1,\,\nu_2)$ is the incomplete gamma function,

$$\Gamma(\nu_1,\,\nu_2) = \int_{\nu_2}^\infty t^{\nu_1-1}e^{-t}\,dt,$$

which goes to zero as ν_2 goes to infinity.

Taking limits in 30 as $z\to\infty$ gives

$$\begin{aligned}\frac{m^a\,p^{-(a+k)}\,\Gamma(a+k)}{(1+\epsilon/(m-\epsilon))^{a+1}} \;&\le\; \lim_{z\to\infty} m^a\int_0^{\frac{\epsilon}{m-\epsilon}} z^{a+k}\frac{t^{k+a-1}}{(1+t)^{a+1}}\,e^{-t\cdot p\cdot z}\,dt\\ &\le m^a\,p^{-(a+k)}\,\Gamma(a+k),\end{aligned}$$

The result follows from (28) the fact that the upper and lower bound are equal as ϵ goes to 0.

□

Continuing with the proof of Proposition 1, we remove the subindex i for simplicity in notation. Since the multivariate Student density can be written as

$$\mathcal{S}t_k(\boldsymbol{\beta} \mid \mathbf{0},\, \boldsymbol{C}^*,\, 2a) = \frac{\Gamma(a+k/2)}{\Gamma(a)}(2\pi)^{-k/2}\, \Big((a\Gamma(a))^{1/a}\sigma^2\rho\,(b+n)\Big)^{-k/2} \\ |\boldsymbol{V}^t\boldsymbol{V}|^{1/2}\, \left(1+\Big(2(a\Gamma(a))^{1/a}\rho\sigma^2(b+n)\Big)^{-1}\,\|\beta\|^2\right)^{-(a+k/2)},$$

it can be easily shown that:

$$\lim_{\|\beta\|^2\to\infty} \frac{\mathcal{S}t_k(\boldsymbol{\beta} \mid \mathbf{0},\, \boldsymbol{C}^*,\, 2a)}{\Gamma(a+k/2)(2\pi)^{-k/2}\,a\left(\sigma^2\rho\,(b+n)\right)^a\,|\boldsymbol{V}^t\boldsymbol{V}|^{1/2}\,2^{a+k/2}\left(\|\boldsymbol{\beta}\|^2\right)^{-(a+k/2)}} =$$

$$= \left(2(a\Gamma(a))^{1/a}\rho\sigma^2(b+n)\lim_{\|\beta\|^2\to\infty}\frac{1+\left(2(a\Gamma(a))^{1/a}\rho\sigma^2(b+n)\right)^{-1}\|\beta\|^2}{\|\beta\|^2}\right)^{-(a+k/2)} = 1.$$

It then follows that

$$\lim_{\|\beta\|^2\to\infty}\frac{\pi^R(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0,\, \sigma)}{\mathcal{S}t_k(\boldsymbol{\beta} \mid \mathbf{0},\, \boldsymbol{C}^*,\, 2a)} = \frac{(2\sigma^2)^{-(a+k/2)}\,b^{-k/2}}{\Gamma(a+k/2)\;(\rho(b+n))^a} \\ \cdot \lim_{\|\beta\|^2\to\infty}(\|\boldsymbol{\beta}\|^2)^{a+k/2}\int_0^1 \lambda^{a-1}\Big(\frac{\lambda}{m-\lambda}\Big)^{k/2} e^{-\frac{\lambda}{m-\lambda}\,p\,\|\beta\|^2}\,d\lambda,$$

where $m = (\rho\,(b+n))/b$ and $p = 1/(2\sigma^2 b)$. Since $\rho > b/(b+n)$ and $m > 1$ we can apply Lemma 2 and the result follows.

A3. Proof of Result 1. To apply invariance, let $\boldsymbol{\theta} = (\boldsymbol{\beta}_0, \sigma, \boldsymbol{\beta}_i, i)$ denote the parameter indexing all the models, and consider the location-scale group defined by $g = (c, \boldsymbol{b}) \in G_0 = (0,\infty) \times \mathcal{R}^{k_0}$ acting on $\boldsymbol{y}$ through the transformation $\tilde{\boldsymbol{y}} = c\,\boldsymbol{y} + \boldsymbol{X}_0\boldsymbol{b}$. It can be easily seen that $\tilde{\boldsymbol{y}} \sim f(\cdot \mid \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_0^*, \sigma^*, \boldsymbol{\beta}_i^*, i^*)$ with $\boldsymbol{\beta}_0^* = \boldsymbol{b} + c\boldsymbol{\beta}_0$, $\sigma^* = c\sigma$, $\boldsymbol{\beta}_i^* = c\boldsymbol{\beta}_i$, and $i^* = i$, so that the transformed model has exactly the same structure as the original model. The Invariance-criterion thus says that the prior $\pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}_0, \sigma)$ must

be such that the marginal models in (5) are invariant with respect to the group action, so that (keeping to the notation above)

$$f(\tilde{\boldsymbol{y}} \mid \boldsymbol{\beta}_0^*, \sigma^*, i^*) = \int \mathcal{N}_n(\tilde{\boldsymbol{y}} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0^* + \boldsymbol{X}\boldsymbol{\beta}^*, (\sigma^*)^2\boldsymbol{I})\, \pi_i(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}_0^*, \sigma^*) d\boldsymbol{\beta}^*\,,$$

the fact that $\pi_i(\cdot \mid \cdot, \cdot)$ must have the same functional form as in the original parameterization following from the completeness of $\mathcal{N}_n(\tilde{\boldsymbol{y}} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0^* + \boldsymbol{X}\boldsymbol{\beta}^*, (\sigma^*)^2\boldsymbol{I})$, given that the design matrix is of full rank. But one can also compute $f(\tilde{\boldsymbol{y}} \mid \boldsymbol{\beta}_0^*, \sigma^*, i^*)$ by change of variables from the original density, yielding

$$f(\tilde{\boldsymbol{y}} \mid \boldsymbol{\beta}_0^*, \sigma^*, i^*) = \int \mathcal{N}_n(\tilde{\boldsymbol{y}} \mid \boldsymbol{X}_0\boldsymbol{\beta}_0^* + \boldsymbol{X}\boldsymbol{\beta}^*, (\sigma^*)^2\boldsymbol{I})\, \pi_i(\boldsymbol{\beta}^*/c \mid (\boldsymbol{\beta}_0^* - \boldsymbol{b})/c, \sigma^*/c)\, c^{-k_0}\, d\boldsymbol{\beta}^*\,.$$

Again using the completeness of the normal density, these two expressions can be equal only if

$$\pi_i(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}_0^*, \sigma^*) = \pi_i(\boldsymbol{\beta}^*/c \mid (\boldsymbol{\beta}_0^* - \boldsymbol{b})/c, \sigma^*/c)\, c^{-k_0}\,.$$

This condition is satisfied by the conditional prior in (12).

With respect to the only if part of the proof, note that for the particular transformation in G_0 given by $\boldsymbol{b} = \boldsymbol{\beta}_0$ and $c = \sigma^*$, the above condition becomes

$$\pi_i(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}_0^*, \sigma^*) = \sigma^{-k_0}\, \pi(\boldsymbol{\beta}^*/\sigma^* \mid \mathbf{0},\, 1),$$

proving that being of the form in (12) is also a necessary condition.

A4. Proof of Result 2. With the use of the full conditional for $\boldsymbol{\beta}_i$ associated with this prior, the integrated models can be alternatively expressed as

$$M_i^I : \boldsymbol{Y}^* = \boldsymbol{X}_0\boldsymbol{\beta}_0 + \sigma\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim f_i^I(\boldsymbol{u})$, given by

$$f_i^I(\boldsymbol{u}) = \int \mathcal{N}_n(\boldsymbol{u} \mid \boldsymbol{X}_i\boldsymbol{t}, \boldsymbol{I})\, h_i(\boldsymbol{t})\, d\boldsymbol{t} \;\; (i > 0) \quad \text{and} \quad f_0^I(\boldsymbol{u}) = \mathcal{N}_n(\boldsymbol{u} \mid \mathbf{0}, \boldsymbol{I}).$$

This model selection problem was explicitly studied in Berger, Pericchi and Varshavsky (1998), where it was shown that the minimal sample size associated with the right-Haar prior for $(\boldsymbol{\beta}_0, \sigma)$ is $n_i^* = k_0 + 1$ and that it is

sufficient for exact predictive matching for $f_i^I(\cdot)$ (or, equivalently, $h_i(\cdot)$) to be symmetric about the origin.

A5. Proof of Result 3. It is convenient to work in terms of orthogonal parameters so, for each model M_i, define $\boldsymbol{\gamma} = \boldsymbol{\beta}_0 + (\boldsymbol{X}_0^t\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^t\boldsymbol{X}_i\,\boldsymbol{\beta}_i$; this will be 'common' to all models and orthogonal to $\boldsymbol{\beta}_i$ in each model M_i, which can be written in the new parameterization as $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\gamma} + \boldsymbol{V}_i\boldsymbol{\beta}_i,\, \sigma^2\boldsymbol{I}_n)$. Consider a scale mixture of normals prior of the form

$$\pi(\boldsymbol{\beta}_i \mid \boldsymbol{\gamma},\, \sigma) = \pi(\boldsymbol{\beta}_i \mid \sigma) = \int_0^\infty \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid \boldsymbol{0},\, g\,\sigma^2\boldsymbol{A}_i)\, h(g)\, dg\,.$$

Noting that the right-Haar prior for $(\boldsymbol{\alpha}, \sigma)$ transforms into the same prior $(1/\sigma)$ for $(\boldsymbol{\gamma}, \sigma)$, it follows that the marginal likelihood under model M_i is

$$\begin{aligned} m_i(\boldsymbol{y}) &= \int \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\gamma} + \boldsymbol{V}_i\boldsymbol{\beta}_i,\, \sigma^2\boldsymbol{I}_n)\, \sigma^{-1}\, \pi_i(\boldsymbol{\beta}_i \mid \boldsymbol{\gamma},\, \sigma)\, d(\boldsymbol{\beta}_i,\, \boldsymbol{\gamma},\, \sigma) \\ &= \int_0^\infty \int \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\gamma} + \boldsymbol{V}_i\boldsymbol{\beta}_i,\, \sigma^2\boldsymbol{I}_n)\, \sigma^{-1}\, \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid \boldsymbol{0},\, g\,\sigma^2\boldsymbol{A}_i) \\ &\qquad\qquad h(g)\, d(\boldsymbol{\beta}_i,\, \boldsymbol{\gamma},\, \sigma) dg\,. \end{aligned}$$

Using the fact that $\boldsymbol{y}^t\boldsymbol{V}_i(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}\boldsymbol{V}_i^t\boldsymbol{y} = SSE_0$ for any sample of size $n = k_i + k_0$ and integrating out $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_i$, and σ yields

$$\begin{aligned} m_i(\boldsymbol{y}) = \int_0^\infty |\boldsymbol{X}_0^t\boldsymbol{X}_0|^{-1/2} &\frac{\pi^{-k_i/2}|(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}|^{1/2}}{2|(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1} + g\boldsymbol{A}_i|^{1/2}} \\ &\left(\hat{\boldsymbol{\beta}}_i^t[(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1} + g\boldsymbol{A}_i]^{-1}\hat{\boldsymbol{\beta}}_i\right)^{-k_i/2} \Gamma\left(\frac{k_i}{2}\right) h(g)\, d(g). \end{aligned}$$

For the robust prior, $\boldsymbol{A}_i = (\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$, and it follows that

$$m_i(\boldsymbol{y}) = \frac{1}{2}|\boldsymbol{X}_0^t\boldsymbol{X}_0|^{-1/2}\pi^{-k_i/2}\Gamma\left(\frac{k_i}{2}\right)(SSE_0)^{-k_i/2}\,,$$

which is the same for all models of dimension k_i, establishing that the robust prior is dimension predictive matching for sample sizes $k_0 + k_i$. Furthermore, this last expression equals $m_0(\boldsymbol{y})$ (see Appendix 6), establishing that the robust prior is null predictive matching for samples of size $k_0 + k_i$. (Note that this result would hold for any proper choice of $h(g)$, not just that for the robust prior.)

To see that null predictive matching does not occur if $\boldsymbol{A}_i$ is not a multiple of $(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$, note that the expression to be established for null predictive matching is (eliminating multiplicative constants)

$$0 = \int_0^\infty \left(\frac{|(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}|^{1/2}(\hat{\boldsymbol{\beta}}_i^t[(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1} + g\boldsymbol{A}_i]^{-1}\hat{\boldsymbol{\beta}}_i)^{-k_i/2}}{|(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1} + g\boldsymbol{A}_i|^{1/2}} - (\hat{\boldsymbol{\beta}}_i^t\boldsymbol{V}_i^t\boldsymbol{V}_i\hat{\boldsymbol{\beta}}_i)^{-k_i/2} \right) h(g)\, d(g).$$

Since $(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$ and $\boldsymbol{A}_i$ are positive definite, there is a matrix $\boldsymbol{B}$ such that $\boldsymbol{B}^t(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}\boldsymbol{B} = \boldsymbol{I}$ and $\boldsymbol{B}^t\boldsymbol{A}_i\boldsymbol{B} = \boldsymbol{D}$, with $\boldsymbol{D}$ being a diagonal matrix with diagonal elements d_i. Also defining $\boldsymbol{W} = \boldsymbol{B}^t\hat{\boldsymbol{\beta}}_i$, it follows that the above expression can be written

$$0 = \int_0^\infty \left(\frac{(W^t[\boldsymbol{I} + g\boldsymbol{D}]^{-1}W)^{-k_i/2}}{|\boldsymbol{I} + g\boldsymbol{D}|^{1/2}} - (|W|^2)^{-k_i/2} \right) h(g)\, d(g)\,.$$

Let d_j be the largest diagonal element and choose W to be the unit vector in coordinate j. Then the above expression becomes

$$0 = \int_0^\infty \left(\frac{(1 + gd_j)^{k_i/2}}{\prod_{l=1}^{k_i}(1 + gd_i)^{1/2}} - 1 \right) h(g)\, d(g)\,.$$

But the integrand is clearly greater than 0, unless all d_i are equal which is equivalent to the statement that $\boldsymbol{A}_i$ is a multiple of $(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$.

A6. Computation of the Bayes factor in (16).

PROPOSITION 2. *For any $(a,\, b,\, \rho_i)$ satisfying (11) and $n \geq k_i + k_0$, the prior predictive distribution for $\boldsymbol{y}$ under M_i using the robust prior is:*

$$m_i^R(\boldsymbol{y}) = m_0^R(\boldsymbol{y})\, Q_{i0}^{-\frac{n-k_0}{2}}\, \frac{2a}{k_i + 2a}\, [\rho_i\,(n+b)]^{-\frac{k_i}{2}}\; AP_{i0},$$

where

$$m_0^R(\boldsymbol{y}) = \frac{1}{2}\, \pi^{-\frac{n-k_0}{2}}\, |\boldsymbol{X}_0^t\boldsymbol{X}_0|^{-\frac{1}{2}}\, \Gamma\left[\frac{n-k_0}{2}\right]\; SSE_0^{-\frac{n-k_0}{2}}$$

and AP_i defined in (3.4.1). Hence the Bayes factor obtained with prior π_i^R in (8) *can be compactly expressed as in (16).*

PROOF. It is convenient to carry out the proof in the orthogonal transformation of the parameters as in Appendix 5. Using standard normal computations, the prior predictive distribution under M_0 is

$$m_0^R(\boldsymbol{y}) = \int_{\mathbb{R}^{k_0}} \int_0^\infty \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\gamma},\, \sigma^2\boldsymbol{I}_n)\, \frac{1}{\sigma}\, d\boldsymbol{\gamma}\, d\sigma$$
$$= \frac{1}{2}\pi^{-\frac{n-k_0}{2}}\, |\boldsymbol{X}_0^t\boldsymbol{X}_0|^{-\frac{1}{2}}\, \Gamma\left[\frac{n-k_0}{2}\right] \mathrm{SSE}_0^{\,-\frac{n-k_0}{2}}\,.$$

Integrating out $\boldsymbol{\beta}_i$, $\boldsymbol{\gamma}$ and σ, the prior predictive distribution under M_i is

$$m_i^R(\boldsymbol{y}) = \int \mathcal{N}_n(\boldsymbol{y} \mid \boldsymbol{X}_0\boldsymbol{\gamma} + \boldsymbol{V}_i\boldsymbol{\beta}_i,\, \sigma^2\boldsymbol{I}_n)\, \mathcal{N}_{k_i}(\boldsymbol{\beta}_i \mid 0, \boldsymbol{B}(\lambda))$$
$$a\, \lambda^{a-1}\, \sigma^{-1}\, d(\boldsymbol{\gamma},\, \boldsymbol{\beta}_i,\, \sigma,\, \lambda) = \frac{1}{2}\pi^{-\frac{n-k_0}{2}}\, |\boldsymbol{X}_0^t\boldsymbol{X}_0|^{-\frac{1}{2}}\, \Gamma\left[\frac{n-k_0}{2}\right]$$
$$\times \int_0^1 a\lambda^{a+\frac{k_i}{2}-1}(\rho_i(b+n) - (b-1)\lambda)^{\frac{n-k_i-k_0}{2}}$$
$$(\mathrm{SSE}_i(\rho_i(b+n) - b\lambda) + \lambda\mathrm{SSE}_0)^{-\frac{n-k_0}{2}}\, d\lambda\,,$$

with $\boldsymbol{B}(\lambda) = (\lambda^{-1}\rho_i(b+n) - b)\sigma^2(\boldsymbol{V}_i^t\boldsymbol{V}_i)^{-1}$. This expression can be rewritten as

$$m_i^R(\boldsymbol{y}) = a\, Q_{i0}^{-\frac{n-k_0}{2}}\, (\rho_i\,(n+b))^{-k_i/2}\, m_0^R(\boldsymbol{y})$$
$$\times \int_0^1 \lambda^{a+\frac{k_i}{2}-1} \left(1 - \frac{b-1}{\rho_i(b+n)}\lambda\right)^{\frac{n-k_i-k_0}{2}} \left(1 - \frac{b - Q_{i0}^{-1}}{\rho_i(b+n)}\lambda\right)^{-\frac{n-k_0}{2}} d\lambda\,,$$

and the result follows by noting that

$$AP_i = \frac{2a+k_i}{2} \int_0^1 \lambda^{a+\frac{k_i}{2}-1} \left(1 - \frac{b-1}{\rho_i(b+n)}\lambda\right)^{\frac{n-k_i-k_0}{2}} \left(1 - \frac{b - Q_{i0}^{-1}}{\rho_i(b+n)}\lambda\right)^{-\frac{n-k_0}{2}} d\lambda\,.$$

□

A7. Proof of Corollary 1.

PROOF. For the prior in (8),

$$\int_0^\infty (1+g)^{-k_i/2}\, p_i^R(g)\, dg = \int_{\rho_i(b+n)-b}^\infty (1+g)^{-k_i/2}\, \frac{a\,[\rho_i(b+n)]^a}{(g+b)^{(a+1)}}\, dg.$$

The change of variables $z = g - [\rho_i(b+n) - b]$ results in

$$\int_0^\infty (1+g)^{-k_i/2}\, p_i^R(g)\, dg = \int_0^\infty \frac{a[\rho_i(b+n)]^a}{[z+\rho_i(b+n)]^{(a+1)}\,[1+z+\rho_i(b+n)-b]^{k_i/2}}\, dz.$$

It is now easy to see that, if $\rho_i(b+n)$ goes to ∞ with n, this integral vanishes as $n \to \infty$ satisfying the condition of Result 5 . □

A8. Proof of Result 7. For simplicity, the explicit dependence of Q_{i0} on $\boldsymbol{y}_m$ will not be shown in this proof, and $\lim_{m\to\infty} Q_{i0}(\boldsymbol{y}_m) = 0$ will be denoted by $Q_{i0} \to 0$. The Robust Bayes factor can be written as

$$\begin{aligned} B_{i0}^R &= a\,(\rho_i(n+b))^{-\frac{k_i}{2}} (Q_{i0})^{-\frac{n-k_0}{2}} \int_0^1 \lambda^{a+\frac{k_i}{2}-1} \left[1 - \frac{b-1}{\rho_i(b+n)}\lambda\right]^{\frac{n-k_i-k_0}{2}} \\ &\qquad \left[1 - \frac{b-Q_{i0}^{-1}}{\rho_i(b+n)}\lambda\right]^{-\frac{n-k_0}{2}} d\lambda \\ &= a\,(\rho_i(n+b))^{-\frac{k_i}{2}} \int_0^1 \lambda^{a+\frac{k_i}{2}-1} \left[1 - \frac{b-1}{\rho_i(b+n)}\lambda\right]^{\frac{n-k_i-k_0}{2}} \\ &\qquad \left[Q_{i0}\left(1 - \frac{b\lambda}{\rho_i(b+n)}\right) + \frac{\lambda}{\rho_i(b+n)}\right]^{-\frac{n-k_0}{2}} d\lambda. \end{aligned}$$

Note that, since $b > 0$, $\rho_i \geq b/(b+n)$, and $0 < \lambda < 1$,

$$\min\{1, \frac{1}{b}\} \leq \left[1 - \frac{b-1}{\rho_i(b+n)}\lambda\right] \leq \max\{1, \frac{1}{b}\}$$

and

$$\left[\frac{\lambda}{\rho_i(b+n)}\right] \leq \left[Q_{i0}\left(1 - \frac{b\lambda}{\rho_i(b+n)}\right) + \frac{\lambda}{\rho_i(b+n)}\right] \leq \left[Q_{i0} + \frac{\lambda}{\rho_i(b+n)}\right].$$

Applying these bounds, it is immediate that

(31)

$$c_1 \int_0^1 \lambda^{a+\frac{k_i}{2}-1} \left[c_2 Q_{i0} + \lambda\right]^{-\frac{n-k_0}{2}}\, d\lambda \leq B_{i0}^R \leq c_3 \int_0^1 \lambda^{a+\frac{k_i}{2}-1} \left[\lambda\right]^{-\frac{n-k_0}{2}}\, d\lambda\,,$$

for positive constants c_1, c_2, and c_3.

To prove the "only if" part of the proposition, note that the last integral in (31) is finite if $n < k_i + k_0 + 2a$. Hence B_{i0} is bounded by a constant as $Q_{i0} \to 0$, and information consistency does not hold.

To prove the "if" part of the proposition, make the change of variables $\lambda^* = \lambda/Q_{i0}$ in the lower bound in (31), resulting in the expression

$$Q_{i0}^{(2a+k_0+k_i-n)/2} c_1 \int_0^{Q_{i0}^{-1}} (\lambda^*)^{a+\frac{k_i}{2}-1} \left[c_2+\lambda^*\right]^{-\frac{n-k_0}{2}} \, d\lambda^* \, .$$

If $n > k_i + k_0 + 2a$, it is clear that this expression goes to infinity as $Q_{i0} \to 0$ (since the integral itself cannot go to 0). If $n = k_i + k_0 + 2a$, the expression becomes

$$c_1 \int_0^{Q_{i0}^{-1}} \left(\frac{\lambda^*}{c_2+\lambda^*}\right)^{a+\frac{k_i}{2}} (\lambda^*)^{-1} \, d\lambda^* \, ,$$

which clearly goes to infinity as as $Q_{i0} \to 0$, completing the proof.